greatlearning
Power Ahead

SMDM PROJECT

Pooja Kabadi
PGP-DSBA Online
Batch- A4
10-10-2021

Table of Contents:

Wholesale Customers Analysis

A wholesale distributor operating in different regions of Portugal has information on annual spending of several items in their stores across different regions and channels. The data consists of 440 large retailers' annual spending on 6 different varieties of products in 3 different regions (Lisbon, Oporto, Other) and across different sales channel (Hotel, Retail).

Exploratory Data Analysis:

	Buyer/Spender	Channel	Region	Fresh	Milk	Grocery	Frozen	Detergents_Paper	Delicatessen
0	1	Retail	Other	12669	9656	7561	214	2674	1338
1	2	Retail	Other	7057	9810	9568	1762	3293	1776
2	3	Retail	Other	6353	8808	7684	2405	3516	7844
3	4	Hotel	Other	13265	1196	4221	6404	507	1788
4	5	Retail	Other	22615	5410	7198	3915	1777	5185
5	6	Retail	Other	9413	8259	5126	666	1795	1451
6	7	Retail	Other	12126	3199	6975	480	3140	545
7	8	Retail	Other	7579	4956	9426	1669	3321	2566
8	9	Hotel	Other	5963	3648	6192	425	1716	750
9	10	Retail	Other	6006	11093	18881	1159	7425	2098

	Buyer/Spender	Channel	Region	Fresh	Milk	Grocery	Frozen	Detergents_Paper	Delicatessen
430	431	Hotel	Other	3097	4230	16483	575	241	2080
431	432	Hotel	Other	8533	5506	5160	13486	1377	1498
432	433	Hotel	Other	21117	1162	4754	269	1328	395
433	434	Hotel	Other	1982	3218	1493	1541	356	1449
434	435	Hotel	Other	16731	3922	7994	688	2371	838
435	436	Hotel	Other	29703	12051	16027	13135	182	2204
436	437	Hotel	Other	39228	1431	764	4510	93	2346
437	438	Retail	Other	14531	15488	30243	437	14841	1867
438	439	Hotel	Other	10290	1981	2232	1038	168	2125
439	440	Hotel	Other	2787	1698	2510	65	477	52

The Dataset consists of 9 Variables

- Channel and Region are categorical columns
- Fresh, Milk, Grocery, Frozen, Detergents Paper, Delicatessen is integer data type.

Checking for Null-values:

```
Data columns (total 9 columns):
 #   Column            Non-Null Count  Dtype
---  ------            --------------  -----
 0   Buyer/Spender     440 non-null    int64
 1   Channel           440 non-null    object
 2   Region            440 non-null    object
 3   Fresh             440 non-null    int64
 4   Milk              440 non-null    int64
 5   Grocery           440 non-null    int64
 6   Frozen            440 non-null    int64
 7   Detergents_Paper  440 non-null    int64
 8   Delicatessen      440 non-null    int64
dtypes: int64(7), object(2)
memory usage: 31.1+ KB
```

There are no null values in the Dataset

1.1.1 Use methods of descriptive statistics to summarize data. Which Region and which Channel spent the most? Which Region and which Channel spent the least?

Descriptive Data Analysis

	count	unique	top	freq	mean	std	min	25%	50%	75%	max
Buyer/Spender	440.0	NaN	NaN	NaN	220.5	127.161315	1.0	110.75	220.5	330.25	440.0
Channel	440	2	Hotel	298	NaN	NaN	NaN	NaN	NaN	NaN	NaN
Region	440	3	Other	316	NaN	NaN	NaN	NaN	NaN	NaN	NaN
Fresh	440.0	NaN	NaN	NaN	12000.297727	12647.328865	3.0	3127.75	8504.0	16933.75	112151.0
Milk	440.0	NaN	NaN	NaN	5796.265909	7380.377175	55.0	1533.0	3627.0	7190.25	73498.0
Grocery	440.0	NaN	NaN	NaN	7951.277273	9503.162829	3.0	2153.0	4755.5	10655.75	92780.0
Frozen	440.0	NaN	NaN	NaN	3071.931818	4854.673333	25.0	742.25	1526.0	3554.25	60869.0
Detergents_Paper	440.0	NaN	NaN	NaN	2881.493182	4767.854448	3.0	256.75	816.5	3922.0	40827.0
Delicatessen	440.0	NaN	NaN	NaN	1524.870455	2820.105937	3.0	408.25	965.5	1820.25	47943.0

Summary of the Dataset:

- The Dataset consists of 2 categorical variables and 6 continuous variables
- The Categorical variables are - 'Channel' and 'Region'
- The Continuous variables are -'Fresh', 'Milk', 'Frozen', 'Grocery', 'Detergents Paper', 'Delicatessen'
- There are 3 unique values in region column in which 'other' has the most entries.
- There are 2 unique values in Channel Column in which 'Hotel' has more entries.
- For all the 6 items, their standard deviation Is more than the mean.
- All the 6 items have significant number of outliers which can be observed based on maximum and 75% Quartile value.

Continuous Variables summary:

1) Fresh

- **Mean:** 12000.29
- **Standard Deviation**: 12647.32
- **Minimum value:** 3.0
- **Maximum Value:** 112151
- **Range:** Max - Min = 112151 - 3 = 112148
- **1st Quartile (25%):** 3127.75
- **2nd Quartile (50%):** 8504.0
- **3rd Quartile (75%):** 16933.75
- **IQR** = Q3 - Q1 = 16933.75 - 3127.75 = 13806.25

2) Milk

- **Mean:** 5796.26
- **Standard Deviation**: 7380.37
- **Minimum value:** 55.0
- **Maximum Value:** 73498.0
- **Range:** Max - Min = 73498.0 – 55.0 = 73443.0
- **1st Quartile (25%):** 1533.0
- **2nd Quartile (50%):** 3627.0
- **3rd Quartile (75%):** 7190.25
- **IQR** = Q3 - Q1 = 7190.25 – 1533.0 = 5657.25

3) Grocery

- **Mean:** 7951.27
- **Standard Deviation**: 9503.16
- **Minimum value:** 3.0
- **Maximum Value:** 92780.0
- **Range:** Max - Min = 92780.0 – 3.0 = 92777.0
- **1st Quartile (25%):** 2153.0
- **2nd Quartile (50%):** 4755.5
- **3rd Quartile (75%):** 10655.75
- **IQR** = Q3 - Q1 = 10655.75 – 2153.0 = 8502.75

4) Frozen

- **Mean:** 3071.93
- **Standard Deviation**: 4854.67
- **Minimum value:** 25.0
- **Maximum Value:** 60869.0
- **Range:** Max - Min = 60869.0 – 25.0= 60844.0
- **1st Quartile (25%):** 742.25
- **2nd Quartile (50%):** 1526.0
- **3rd Quartile (75%):** 3554.25
- **IQR** = Q3 - Q1 = 3554.25 – 742.25 = 2812.0

5) Detergents Paper

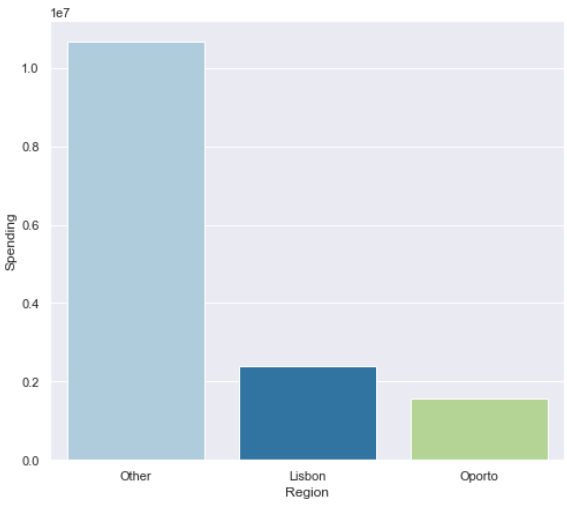
- **Mean:** 2881.49
- **Standard Deviation**: 4767.85
- **Minimum value:** 3.0
- **Maximum Value:** 40827.0
- **Range:** Max - Min = 40827.0 – 3.0 = 40824.0
- **1st Quartile (25%):** 256.75
- **2nd Quartile (50%):** 816.5
- **3rd Quartile (75%):** 3922.0
- **IQR** = Q3 - Q1 = 3922.0 – 256.75 = 3665.25

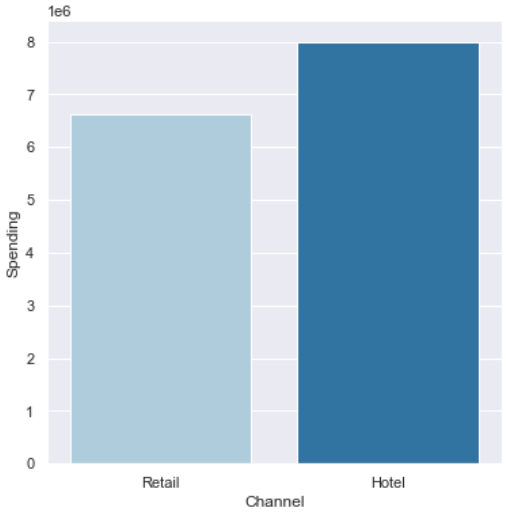
6) Delicatessen

- **Mean:** 1524.87
- **Standard Deviation**: 2820.10
- **Minimum value:** 3.0
- **Maximum Value:** 47943.0
- **Range:** Max - Min = 47943.0 – 3.0 = 47940.0
- **1st Quartile (25%):** 408.25
- **2nd Quartile (50%):** 965.5
- **3rd Quartile (75%):** 1820.25
- **IQR** = Q3 - Q1 = 1820.25 – 408.25 = 1412.0

1.1.2 Which Region and which Channel spent the most?

1.1.3 Which Region and which Channel spent the least?






```
Region
Lisbon     2386813
Oporto     1555088
Other     10677599
Name: Spending, dtype: int64

Channel
Hotel     7999569
Retail    6619931
Name: Spending, dtype: int64
```

- Based on above Graphs, we can see that 'Hotel' channel gets the maximum expenditure and 'retail' channel get the minimum expenditure.
- Also, as per region wise, 'other' region has maximum expenditure and 'Oporto' region has minimum expenditure.

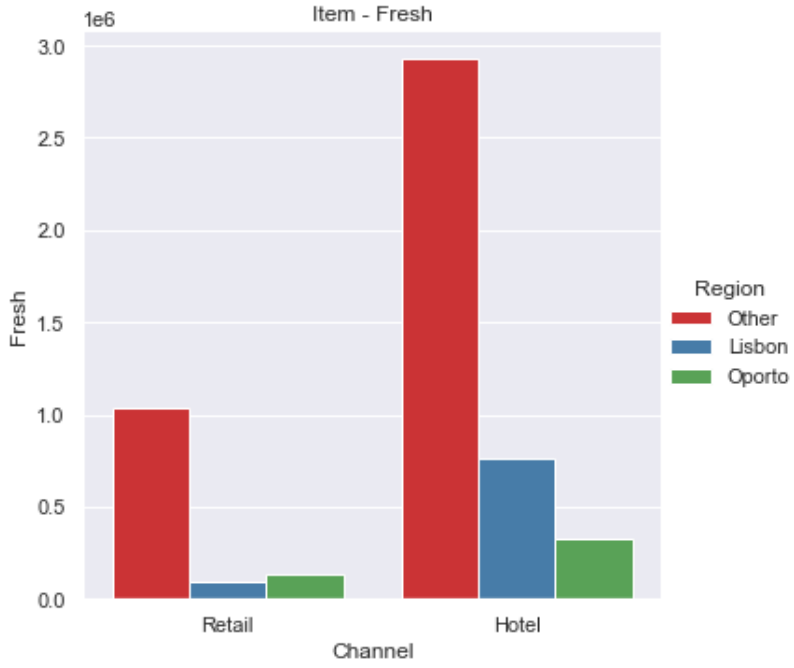
Highest spend in the Region is from "Others (1,06,77,599)" and lowest spend in the region is from "Oporto (15,55,088)"

Highest spend in the Channel is from "Hotel (79,99,569)" and lowest spend in the Channel is from "Retail (66,19,931)"

1.2. There are 6 different varieties of items that are considered. Describe and comment/explain all the varieties across Region and Channel? Provide a detailed justification for your answer.

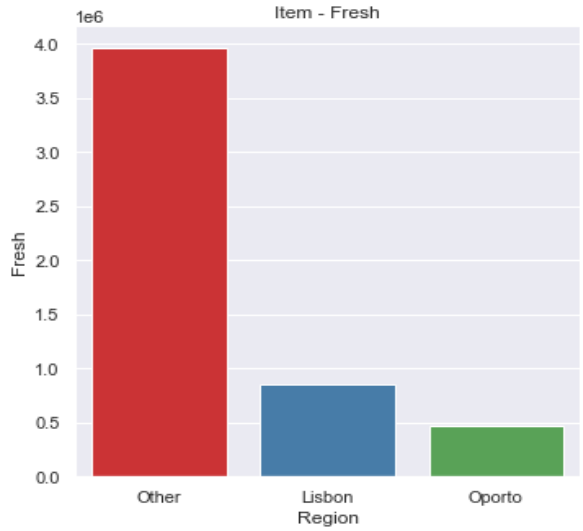
There are 6 categories across Regions and Channels are Fresh, Milk, Grocery, Frozen, Detergents Paper & Delicatessen

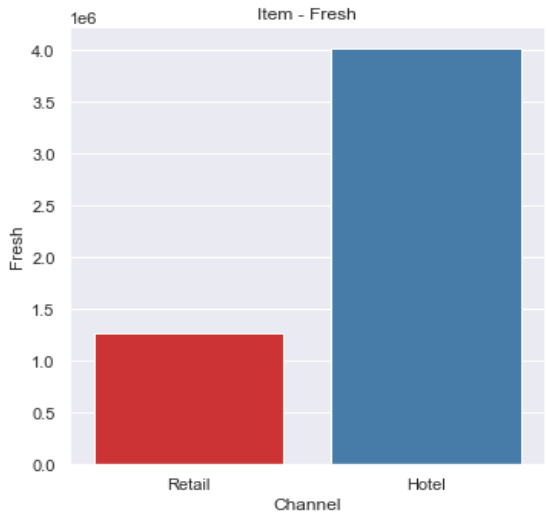
1) Fresh



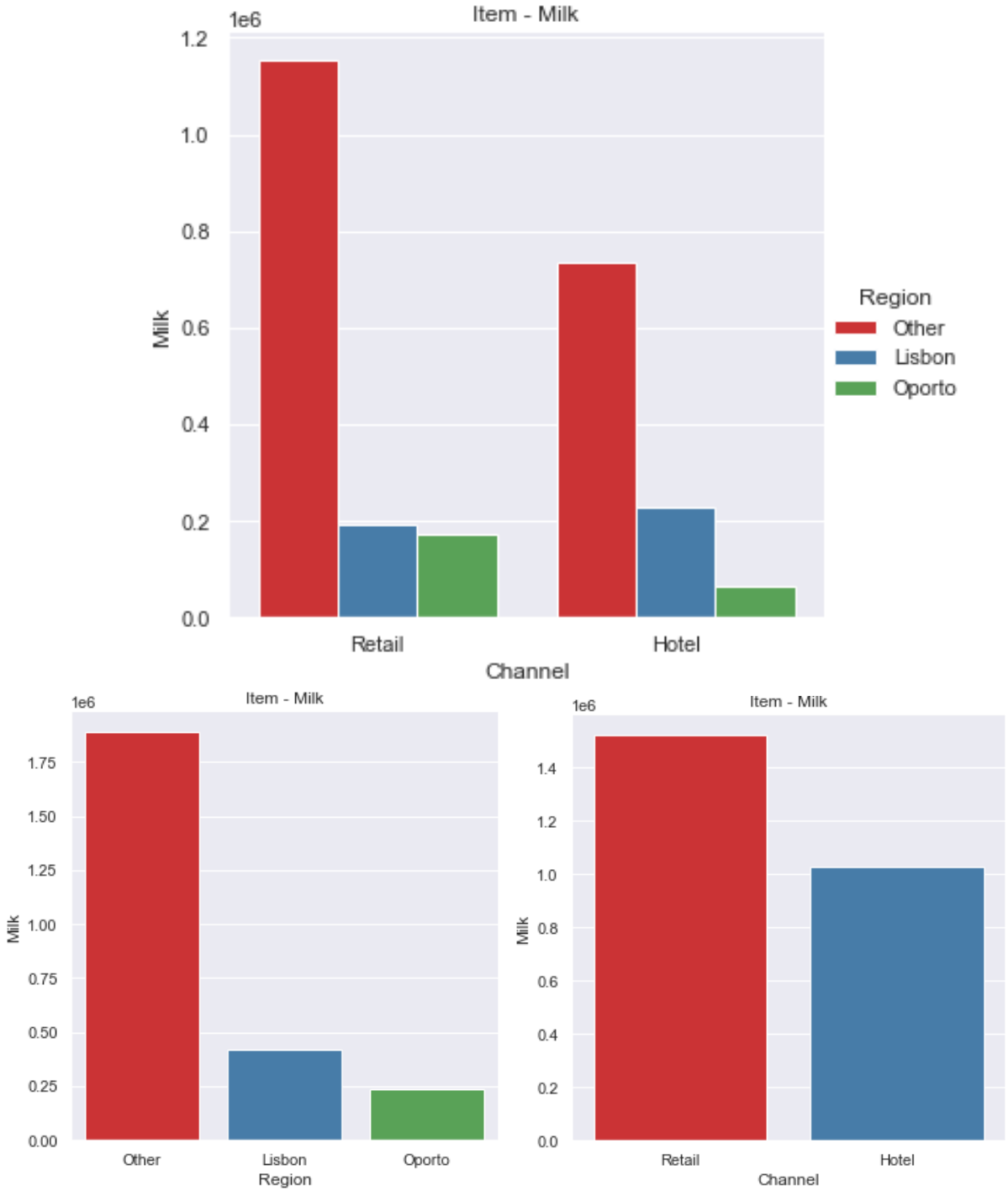
As per the plots, we can see that:

- In 'Retail' channel, Fresh item is most sold in 'Other' region followed by the 'Oporto' and then by the 'Lisbon' region.
- In 'Hotel' channel, Fresh item is most sold in 'Other' region followed by 'Lisbon' and 'Oporto' region.
- Across all the 3 regions, Fresh item is sold most in the 'Hotel' region compared to the 'Retail' region.





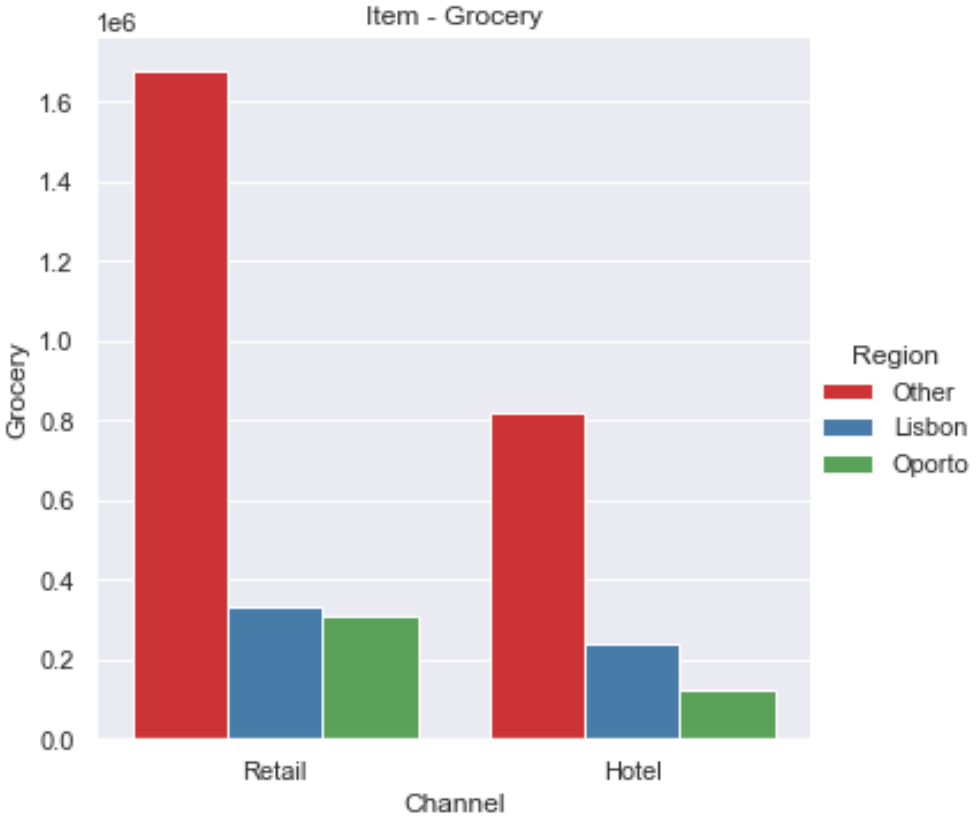
2) Milk

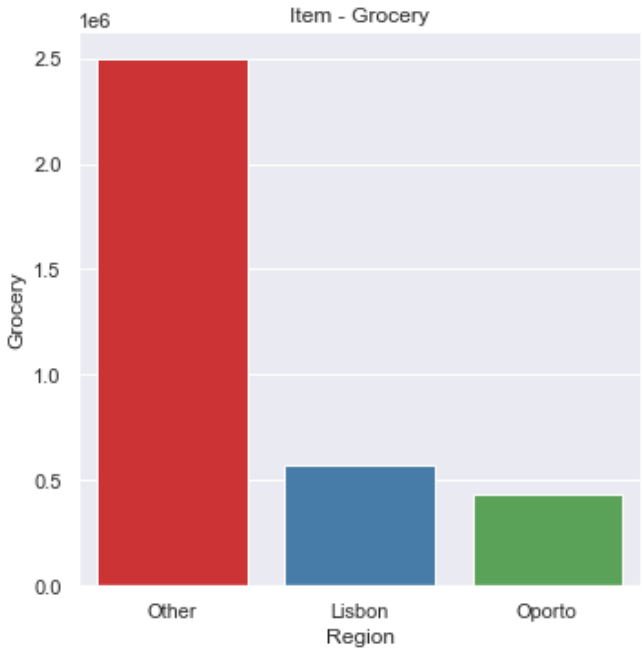


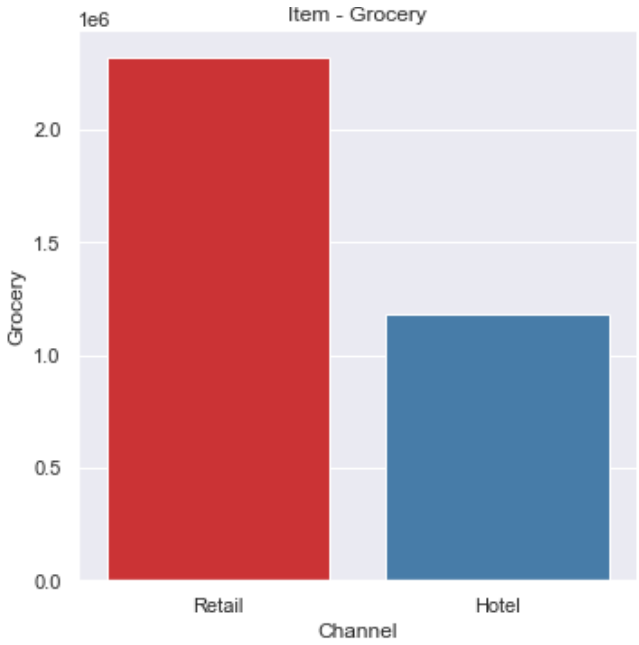
As per the plots, we can see that:

- The Milk item is most sold in 'Retail' channel compared to the 'Hotel' channel.
- The milk item is most sold in 'Other' region followed up by the 'Lisbon' and 'Oporto' region.
- In 'Retail' channel, milk items are most sold in 'Other' region followed by 'Lisbon' and 'Oporto'.
- In 'Hotel' channel milk items are most sold in 'Other' region followed by 'Lisbon' and 'Oporto'.

3) Grocery



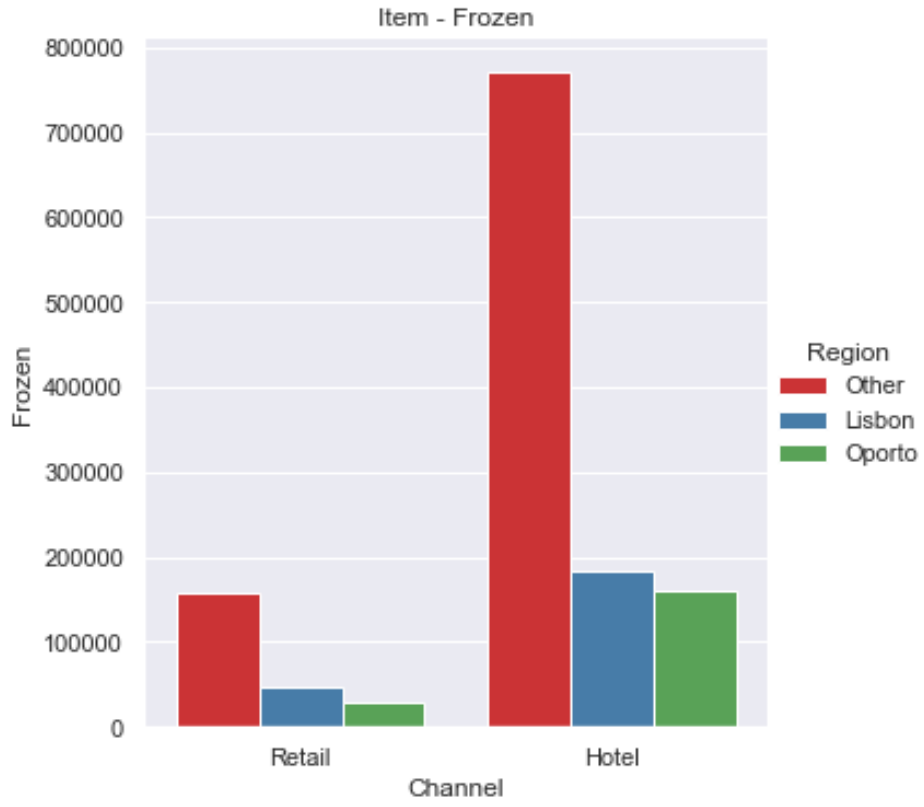


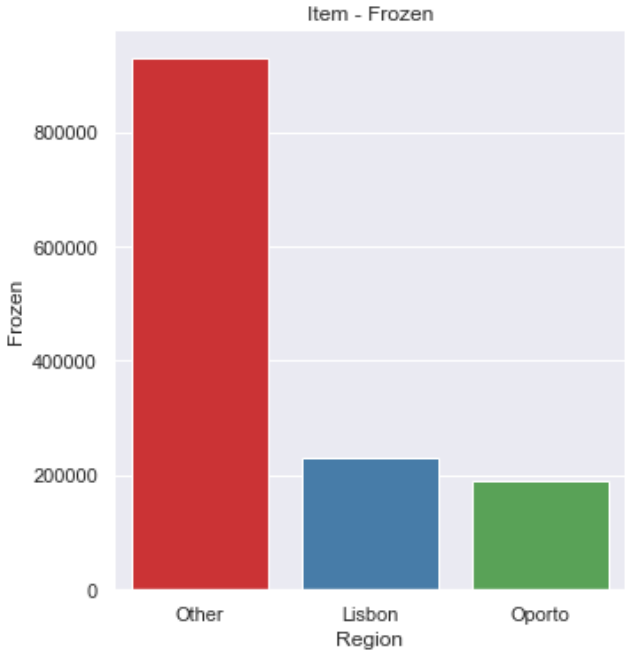


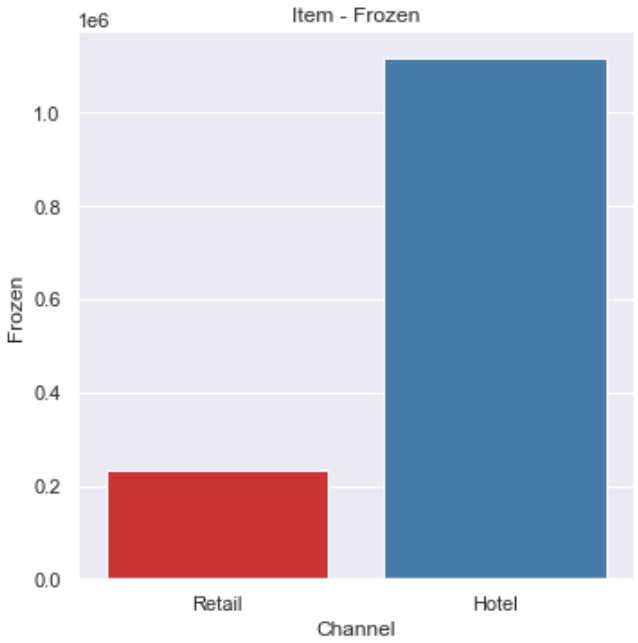
As per the plots, we can see that:

- The Grocery item is most sold in ‘Retail’ channel compared to the ‘Hotel’ channel.
- In all Regions, Grocery items are maximum sold in ‘other’ region followed by ‘Lisbon’ and ‘Oporto’

4) Frozen



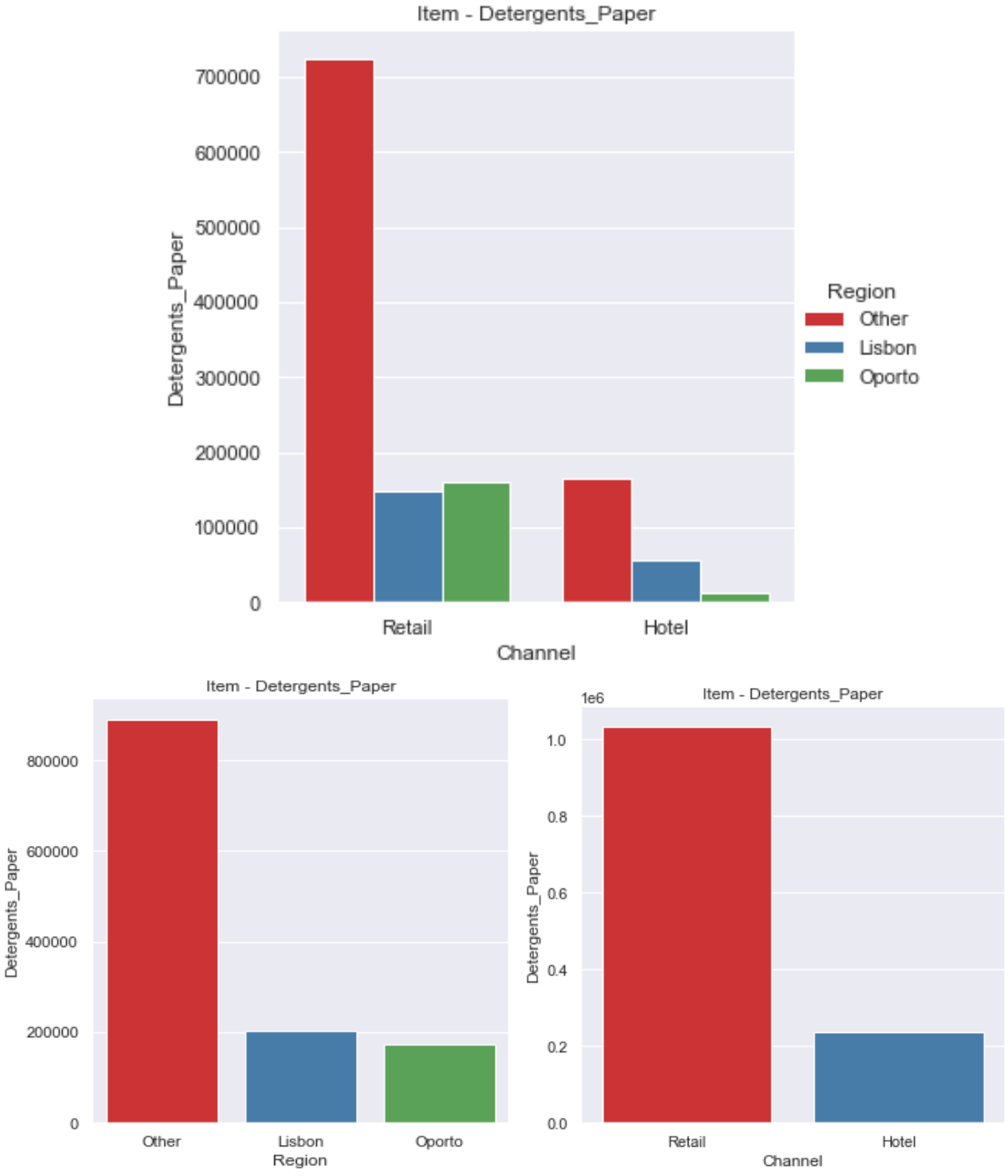




As per the plots, we can see that:

- The Frozen item is most sold in ‘Hotel’ channel compared to the ‘Retail’ channel.
- The Frozen item is most sold in ‘Other’ region followed up by the ‘Lisbon’ and ‘Oporto’ region.
- In both Channels, Frozen items are maximum sold in ‘other’ region followed by ‘Lisbon’ and ‘Oporto’

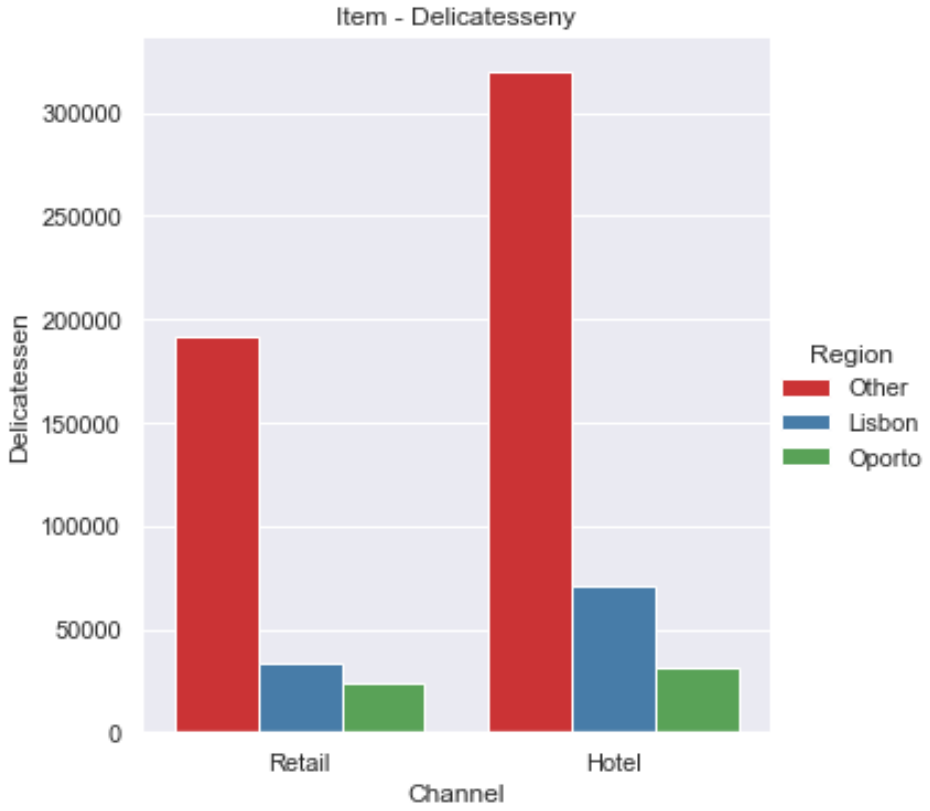
5) Detergents Paper

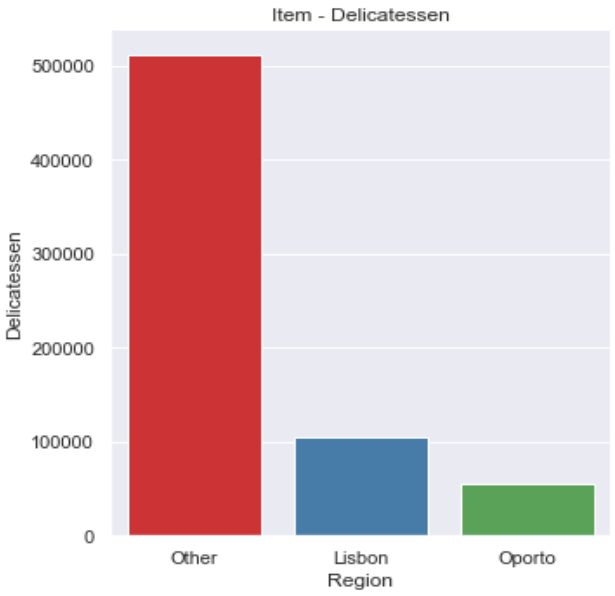


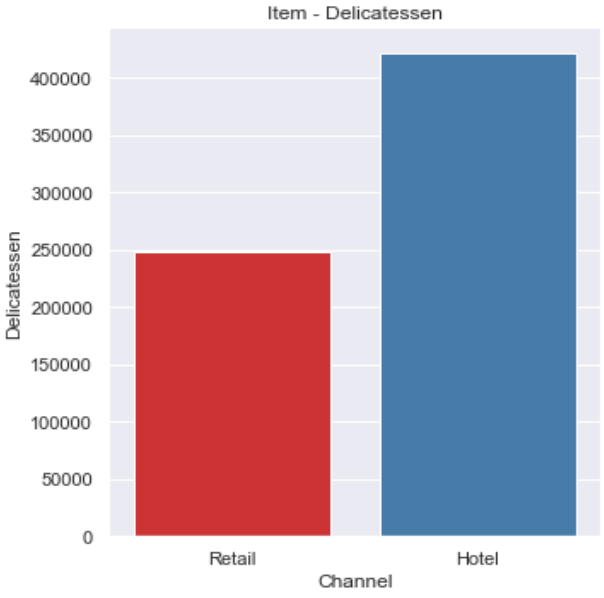
As per the plots, we can see that:

- The Detergents Paper item is most sold in ‘Retail’ channel compared to the ‘Hotel’ channel.
- The Detergents Paper item is most sold in ‘Other’ region followed up by the ‘Lisbon’ and ‘Oporto’ region.
- In ‘Retail’ channel, Detergents Paper items are most sold in ‘Other’ region followed by ‘Oporto’ and ‘Lisbon’.
- In ‘Hotel’ channel Detergents Paper items are most sold in ‘Other’ region followed by ‘Lisbon’ and ‘Oporto’.

6) Delicatessen



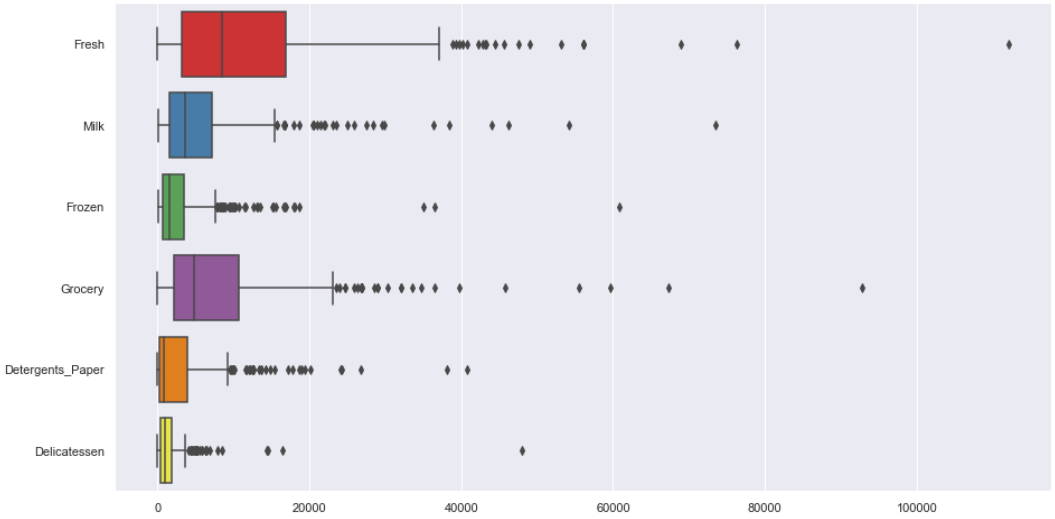




As per the plots, we can see that:

- The Delicatessen item is most sold in ‘Hotel’ channel compared to the ‘Retail’ channel.
- The Delicatessen item is most sold in ‘Other’ region followed up by the ‘Lisbon’ and ‘Oporto’ region.
- In ‘Retail’ channel, Delicatessen items are most sold in ‘Other’ region followed by ‘Lisbon’ and ‘Oporto’.
- In ‘Hotel’ channel Delicatessen items are most sold in ‘Other’ region followed by ‘Lisbon’ and ‘Oporto’.

Checking the spending across each item by using box plot:



we see that the spending for 'Fresh' and 'Grocery' is the highest across Region and Channel and Delicatessen being the least spending.

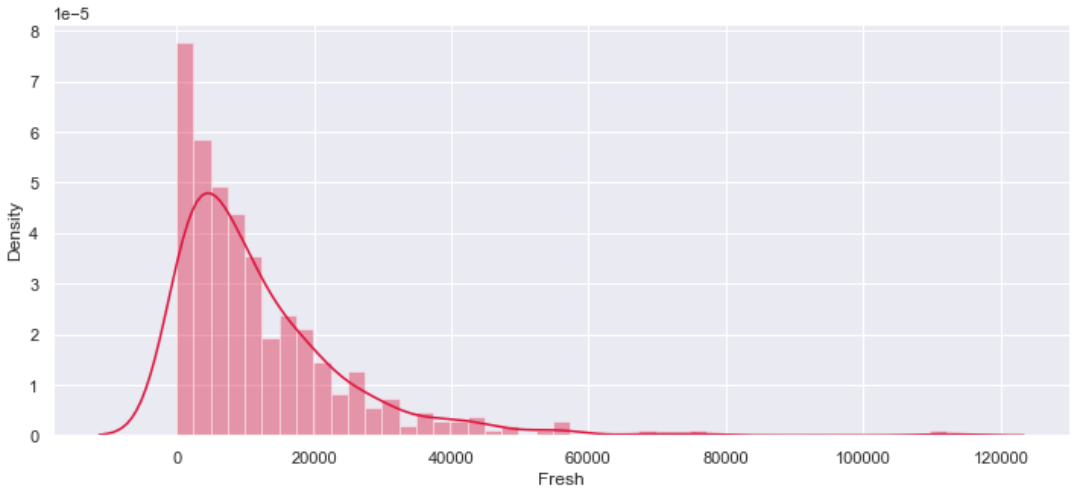
1.3. On the basis of the descriptive measure of variability, which item shows the most inconsistent behaviour? Which items shows the least inconsistent behaviour?

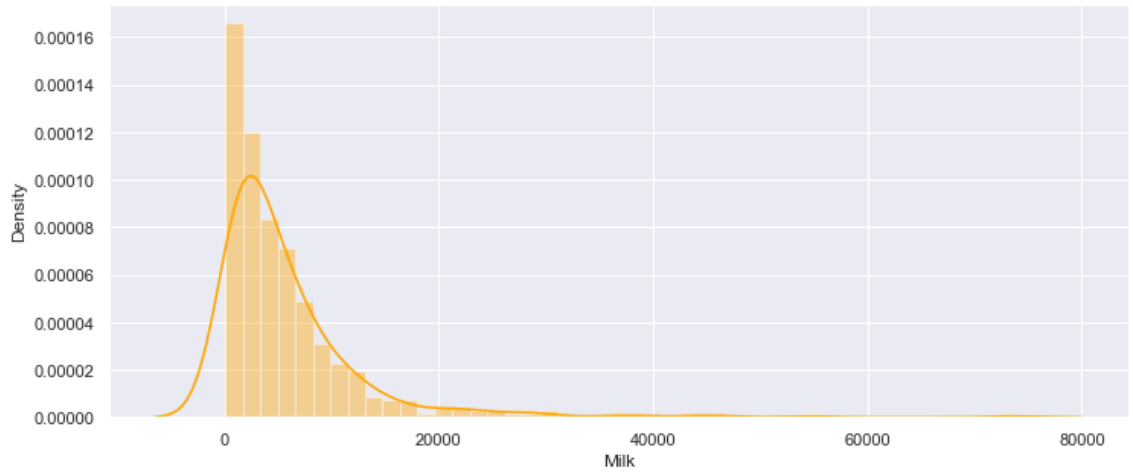
SL.NO.	ITEM	MEAN	STANDARD DEVIATION	COEFFICIENT OF VARIATION
1	Fresh	12000.29	12647.32	1.053
2	Milk	5796.26	7380.37	1.273
3	Frozen	3071.93	4854.67	1.580
4	Grocery	7951.27	9503.16	1.195
5	Detergents Paper	2881.49	4767.85	1.654
6	Delicatessen	1524.87	2820.10	1.849

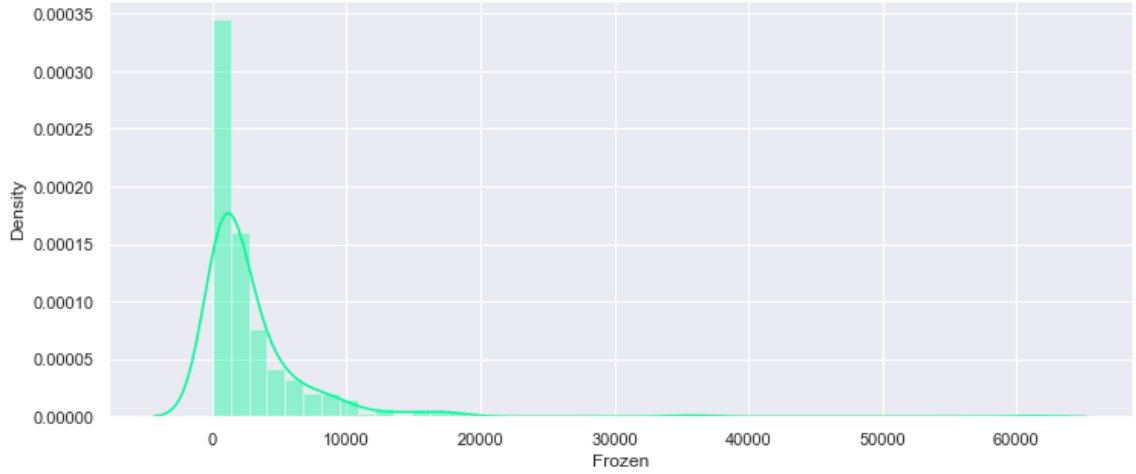
- From the above table we can see that 'Fresh' items have least coefficient of variation and shows least inconsistent

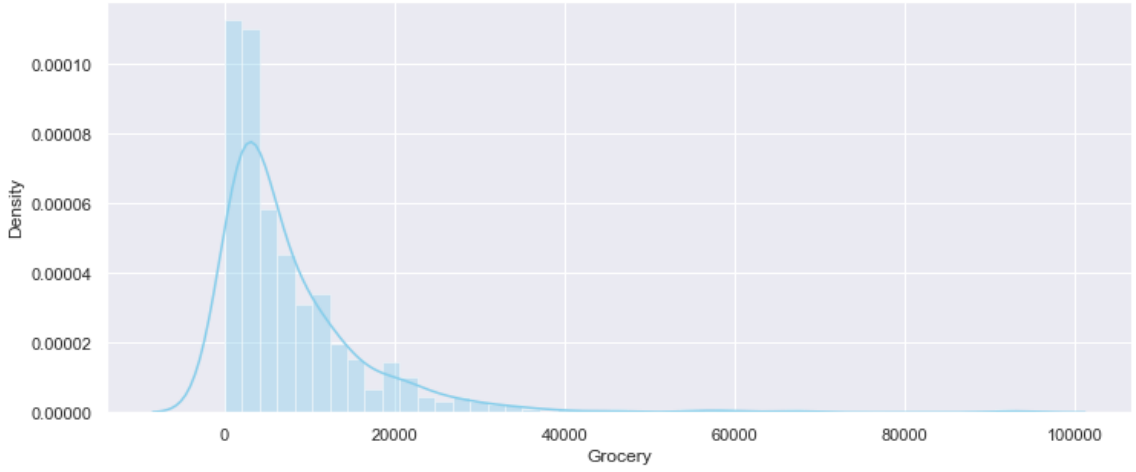
- 'Delicatessen' items have highest coefficient of Variation and shows the most inconsistent behaviour.

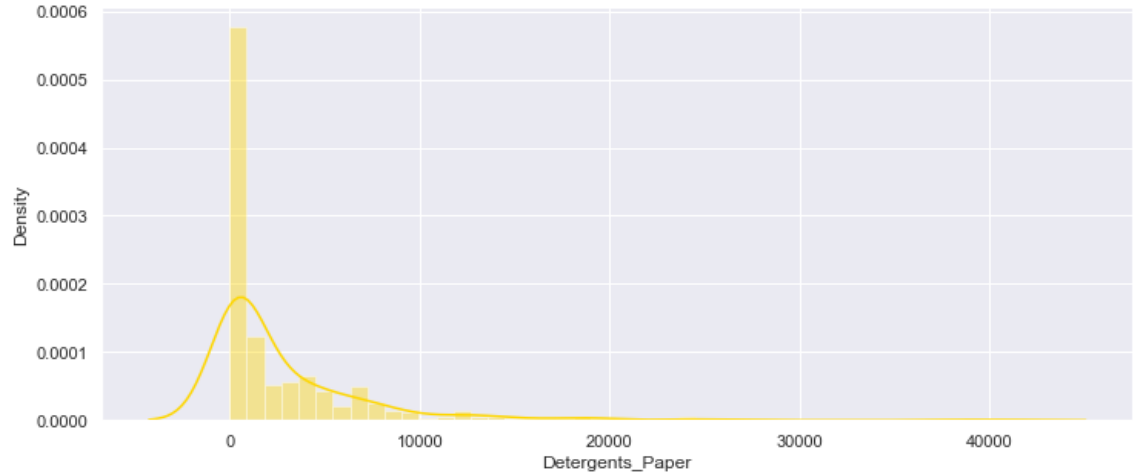
-

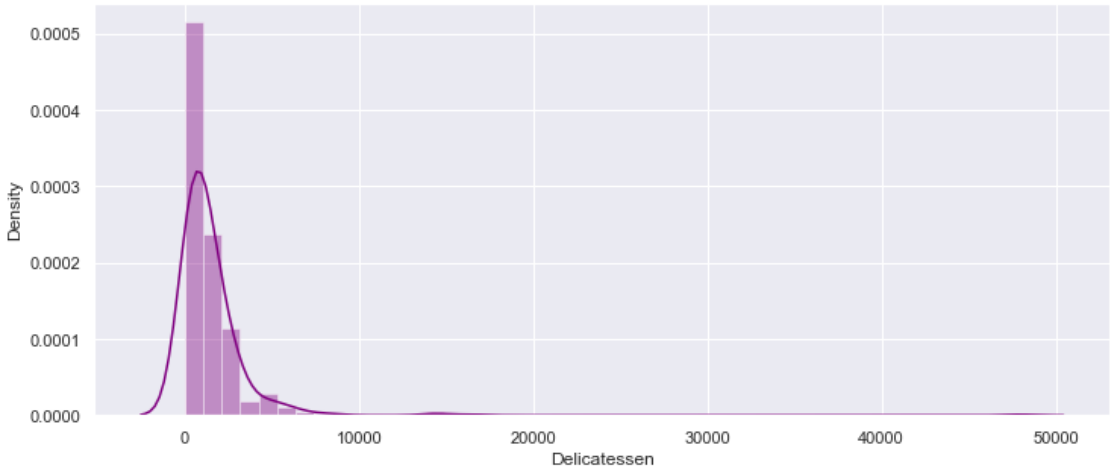



Density
0.00016
0.00014
0.00012
0.00010
0.00008
0.00006
0.00004
0.00002
0.00000
0
20000
40000
60000
80000
Milk


Density
0.00035
0.00030
0.00025
0.00020
0.00015
0.00010
0.00005
0.00000
0
10000
20000
30000
40000
50000
60000
Frozen


Density
0.00010
0.00008
0.00006
0.00004
0.00002
0.00000
0
20000
40000
60000
80000
100000
Grocery


Density
0.0006
0.0005
0.0004
0.0003
0.0002
0.0001
0.0000
0
10000
20000
30000
40000
Detergents_Paper

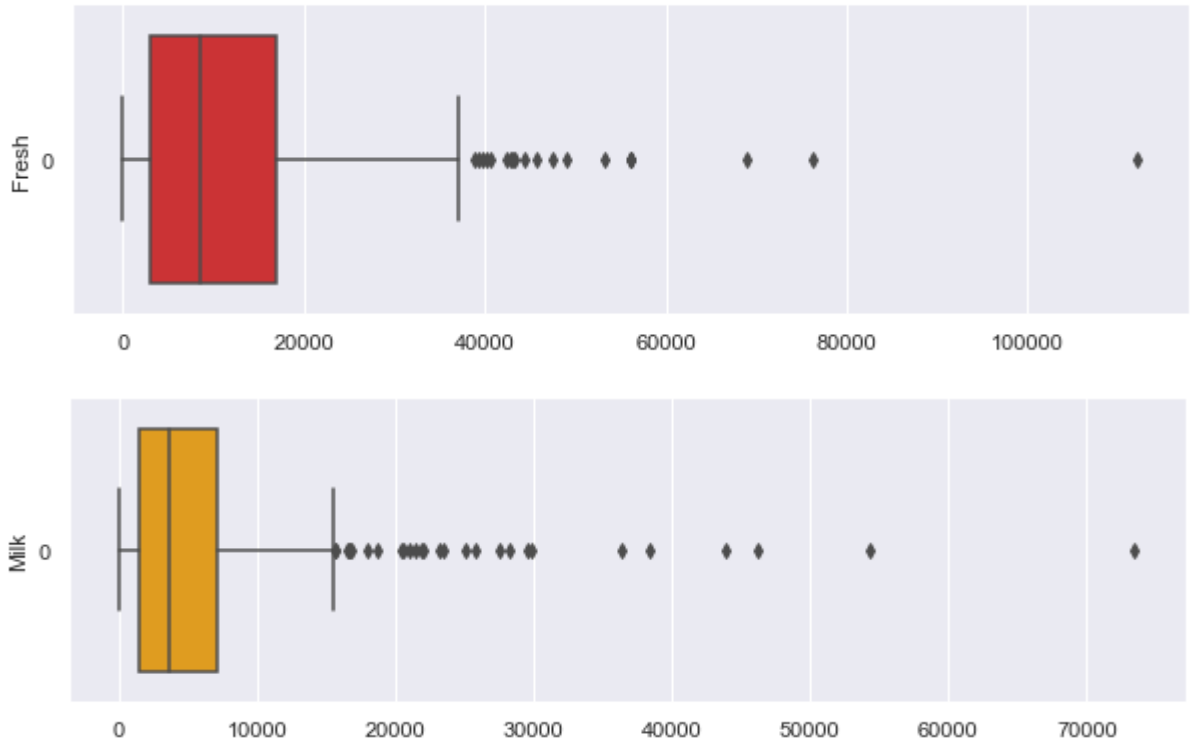


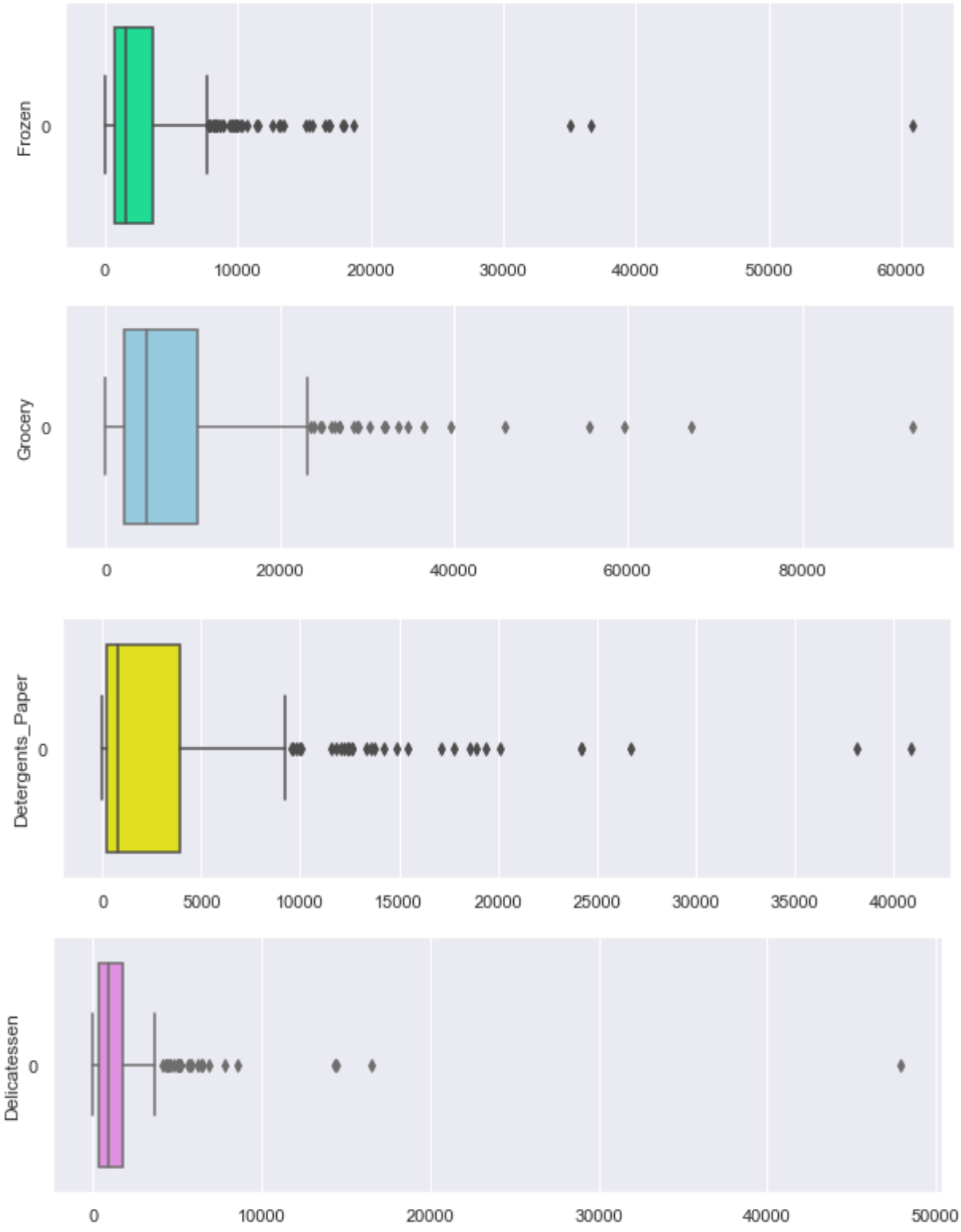
The histograms above of the items, we can show the same that the 'Fresh' items are the most widespread among the items and have the highest standard deviation and whereas 'Delicatessen' items being less variable have the lowest standard deviation.

Least inconsistent – Fresh Item

Most inconsistent – Delicatessen Item.

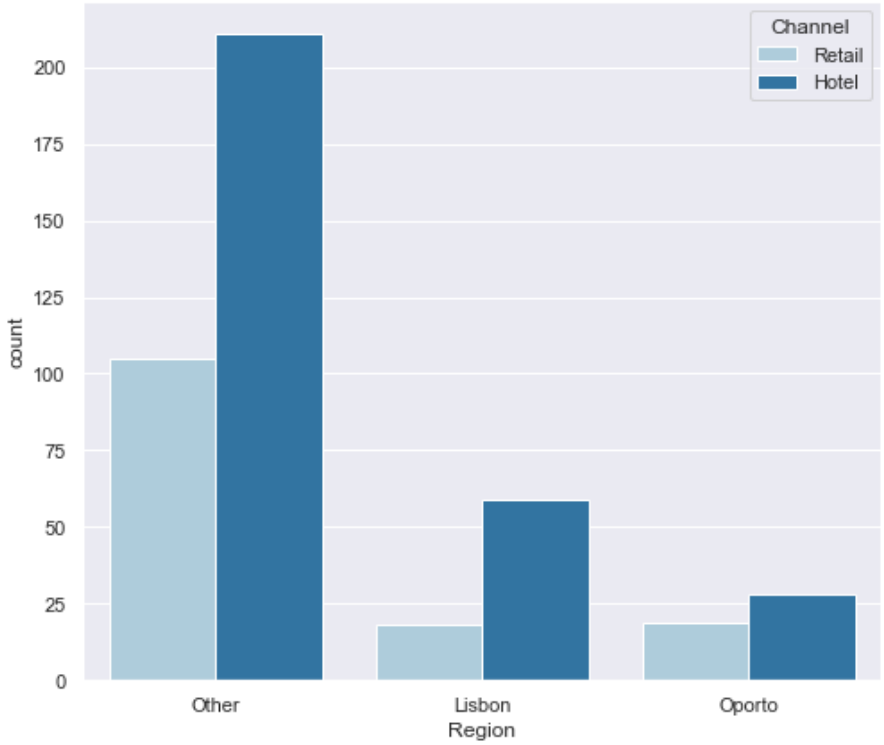
1.4 Are there any outliers in the data? Back up your answer with a suitable plot/technique with the help of detailed comments.





From the above Box plots, we can clearly see that there are outliers in each of the Item in the dataset.

1.5. On the basis of your analysis, what are your recommendations for the business? How can your analysis help the business to solve its problem? Answer from the business perspective.



Business Recommendations:

- The Spending on 'Hotel' and 'Retail' are different, the spending of 'Hotel' channel is more than the 'Retail' channel which should be more or else equal, so the 'Retail' channel should be focused.
- From the above analysis and patterns, we can see that the most of the spending is from the 'Fresh' item followed by the 'Grocery' item. The other items should be focused as well.
- Fresh items are more in demand in the 'Hotel' channel in 'Other' region.
- Grocery items are more in demand in Retail channel.
- Delicatessen and Detergent papers are in very less demand in all the regions, therefore expenditure for items should be given accordingly.
- We can also see from the above box plots that, all the items have many outliers, the total spending are greater than mean spending. The outliers should be minimized.

Clear Mountain State University (CMSU) Survey Analysis

The Student News Service at Clear Mountain State University (CMSU) has decided to gather data about the undergraduate students that attend CMSU. CMSU creates and distributes a survey of 14 questions and receives responses from 62 undergraduates (stored in the ***Survey*** data set).

Exploratory Data Analysis:

	ID	Gender	Age	Class	Major	Grad Intention	GPA	Employment	Salary	Social Networking	Satisfaction	Spending	Computer	Text Messages
0	1	Female	20	Junior	Other	Yes	2.9	Full-Time	50.0	1	3	350	Laptop	200
1	2	Male	23	Senior	Management	Yes	3.6	Part-Time	25.0	1	4	360	Laptop	50
2	3	Male	21	Junior	Other	Yes	2.5	Part-Time	45.0	2	4	600	Laptop	200
3	4	Male	21	Junior	CIS	Yes	2.5	Full-Time	40.0	4	6	600	Laptop	250
4	5	Male	23	Senior	Other	Undecided	2.8	Unemployed	40.0	2	4	500	Laptop	100
5	6	Female	22	Senior	Economics/Finance	Undecided	2.3	Unemployed	78.0	3	2	700	Laptop	30
6	7	Female	21	Junior	Other	Undecided	3.0	Part-Time	50.0	1	3	500	Laptop	50
7	8	Female	22	Senior	Other	Undecided	3.1	Full-Time	80.0	1	2	200	Tablet	300
8	9	Female	20	Junior	Management	Yes	3.6	Unemployed	30.0	0	4	500	Laptop	400
9	10	Female	21	Senior	Economics/Finance	Undecided	3.3	Part-Time	37.5	1	4	200	Laptop	100

	ID	Gender	Age	Class	Major	Grad Intention	GPA	Employment	Salary	Social Networking	Satisfaction	Spending	Computer	Text Messages
52	53	Female	21	Senior	Retailing/Marketing	Undecided	3.7	Part-Time	40.0	3	4	300	Laptop	700
53	54	Male	21	Junior	Retailing/Marketing	No	3.4	Part-Time	40.0	1	5	500	Laptop	300
54	55	Male	21	Senior	Other	Yes	3.4	Part-Time	50.0	1	4	250	Desktop	700
55	56	Female	21	Senior	Retailing/Marketing	No	3.1	Part-Time	50.0	1	1	300	Laptop	300
56	57	Female	21	Senior	International Business	Yes	3.4	Part-Time	42.0	1	1	200	Laptop	100
57	58	Female	21	Senior	International Business	No	2.4	Part-Time	40.0	1	3	1000	Laptop	10
58	59	Female	20	Junior	CIS	No	2.9	Part-Time	40.0	2	4	350	Laptop	250
59	60	Female	20	Sophomore	CIS	No	2.5	Part-Time	55.0	1	4	500	Laptop	500
60	61	Female	23	Senior	Accounting	Yes	3.5	Part-Time	30.0	2	3	490	Laptop	50
61	62	Female	23	Senior	Economics/Finance	No	3.2	Part-Time	70.0	2	3	250	Laptop	0

Checking for Null-values:

```
Data columns (total 14 columns):
 #   Column             Non-Null Count  Dtype
---  ------             --------------  -----
 0   ID                 62 non-null     int64
 1   Gender             62 non-null     object
 2   Age                62 non-null     int64
 3   Class              62 non-null     object
 4   Major              62 non-null     object
 5   Grad Intention     62 non-null     object
 6   GPA                62 non-null     float64
 7   Employment         62 non-null     object
 8   Salary             62 non-null     float64
 9   Social Networking  62 non-null     int64
 10  Satisfaction       62 non-null     int64
 11  Spending           62 non-null     int64
 12  Computer           62 non-null     object
 13  Text Messages      62 non-null     int64
dtypes: float64(2), int64(6), object(6)
memory usage: 6.9+ KB
```

Data has 14 variables in it

- There are 6 categorical variables that are Gender, Class, major, Grad Intent, Employment and Computer.
- There are 5 integer data type variables that are Age, Social Networking, Satisfaction, Spending and Text Messages.
- There GPA and Salary are 2 float data type variables

Descriptive Data Analysis

	count	unique	top	freq	mean	std	min	25%	50%	75%	max
ID	62.0	NaN	NaN	NaN	31.5	18.041619	1.0	16.25	31.5	46.75	62.0
Gender	62	2	Female	33	NaN	NaN	NaN	NaN	NaN	NaN	NaN
Age	62.0	NaN	NaN	NaN	21.129032	1.431311	18.0	20.0	21.0	22.0	26.0
Class	62	3	Senior	31	NaN	NaN	NaN	NaN	NaN	NaN	NaN
Major	62	8	Retailing/Marketing	14	NaN	NaN	NaN	NaN	NaN	NaN	NaN
Grad Intention	62	3	Yes	28	NaN	NaN	NaN	NaN	NaN	NaN	NaN
GPA	62.0	NaN	NaN	NaN	3.129032	0.377388	2.3	2.9	3.15	3.4	3.9
Employment	62	3	Part-Time	43	NaN	NaN	NaN	NaN	NaN	NaN	NaN
Salary	62.0	NaN	NaN	NaN	48.548387	12.080912	25.0	40.0	50.0	55.0	80.0
Social Networking	62.0	NaN	NaN	NaN	1.516129	0.844305	0.0	1.0	1.0	2.0	4.0
Satisfaction	62.0	NaN	NaN	NaN	3.741935	1.213793	1.0	3.0	4.0	4.0	6.0
Spending	62.0	NaN	NaN	NaN	482.016129	221.953805	100.0	312.5	500.0	600.0	1400.0
Computer	62	3	Laptop	55	NaN	NaN	NaN	NaN	NaN	NaN	NaN
Text Messages	62.0	NaN	NaN	NaN	246.209677	214.46595	0.0	100.0	200.0	300.0	900.0

2.1. For this data, construct the following contingency tables (Keep Gender as row variable)

2.1.1. Gender and Major

Major Gender	Accounting	CIS	Economics/Finance	International Business	Management	Other	Retailing/Marketing	Undecided
Female	3	3	7	4	4	3	9	0
Male	4	1	4	2	6	4	5	3

2.1.2. Gender and Grad Intention

Grad Intention Gender	No	Undecided	Yes
Female	9	13	11
Male	3	9	17

2.1.3. Gender and Employment

Employment / Gender	Full-Time	Part-Time	Unemployed
Female	3	24	6
Male	7	19	3

2.1.4. Gender and Computer

Computer / Gender	Desktop	Laptop	Tablet
Female	2	29	2
Male	3	26	0

2.2. Assume that the sample is representative of the population of CMSU. Based on the data, answer the following question:

2.2.1. What is the probability that a randomly selected CMSU student will be male?

Number of male students = 29

Total number of students = 62

Probability that a randomly selected CMSU student will be male = 29/62

P(Male) = 0.468 = 46.8%

2.2.2. What is the probability that a randomly selected CMSU student will be female?

Number of male students = 33

Total number of students = 62

Probability that a randomly selected CMSU student will be male = 33/62

P(Female) = 0.532 = 53.2%

2.3. Assume that the sample is representative of the population of CMSU. Based on the data, answer the following question:

2.3.1. Find the conditional probability of different majors among the male students in CMSU.

Total number of students = 62 and Number of males = 29

Major Gender	Accounting	CIS	Economics/Finance	International Business	Management	Other	Retailing/Marketing	Undecided
Female	3	3	7	4	4	3	9	0
Male	4	1	4	2	6	4	5	3

Probability of male opting for accounting = 4/ 29

P (Accounting| Male) = 0.138

Probability of male opting for CIS = 1/ 29

P (CIS| Male) = 0.0344

Probability of male opting for Economics/Finance = 4/ 29

P (Economics/Finance| Male) = 0.138

Probability of male opting for International Business = 2/ 29

P (International Business| Male) = 0.069

Probability of male opting for Management = 6/ 29

P (Management | Male) = 0.207

Probability of male opting for Other = 4/ 29

P (Other | Male) = 0.138

Probability of male opting for Retailing/Marketing = 5/ 29

P (Retailing/Marketing | Male) = 0.172

Probability of male opting for Undecided = 3/ 29

P (Undecided | Male) = 0.103

Results:

- Probability of Males opting for accounting is 13.79%
- Probability of Males opting for CIS. is 3.45%
- Probability of Males opting for Economics/Finance is 13.79%
- Probability of Males opting for International Business is 6.90%
- Probability of Males opting for Management is 20.69%
- Probability of Males opting for Other is 13.79%
- Probability of Males opting for Retailing/Marketing is 17.24%
- Probability of Males of Undecided is 10.34%

2.3.2 Find the conditional probability of different majors among the female students of CMSU.

Total number of students = 62

Number of males = 33

Probability of female opting for accounting = 3/ 33

P (Accounting| Female) = 0.090

Probability of female opting for CIS = 3/ 33

P (CIS| Female) = 0.090

Probability of female opting for Economics/Finance = 7/ 33

P (Economics/Finance| Female) = 0.212

Probability of female opting for International Business = 4/ 33

P (International Business| Female) = 0.121

Probability of female opting for Management = 4/ 33

P (Management | Female) = 0.121

Probability of female opting for Other = 3/ 33

P (Other | Female) = 0.090

Probability of female opting for Retailing/Marketing = 9/ 33

P (Retailing/Marketing | Female) = 0.273

Probability of female opting for Undecided = 0/ 33

P (Undecided | Female) = 0.0

Results:

- Probability of Females opting for accounting is 9.09%
- Probability of Females opting for CIS is 9.09%
- Probability of Females opting for Economics/Finance is 21.21%
- Probability of Females opting for international Business is 12.12%
- Probability of Females opting for Management is 12.12%
- Probability of Females opting for Other is 9.09%
- Probability of Females opting for Retailing/Marketing is 27.27%
- Probability of Females of Undecided is 0.0%

2.4. Assume that the sample is a representative of the population of CMSU. Based on the data, answer the following question:

2.4.1. Find the probability that a randomly chosen student is a male and intends to graduate.

Grad Intention	No	Undecided	Yes
Gender			
Female	9	13	11
Male	3	9	17

Male Student and Intend to Graduate = 17

Total Students = 62

Probability of randomly chosen student is a male and intends to graduate = 17/62 = 0.274

Probability of randomly chosen student is a male and intends to graduate = 27.4%

2.4.2 Find the probability that a randomly selected student is a female and does NOT have a laptop.

Computer	Desktop	Laptop	Tablet
Gender			
Female	2	29	2
Male	3	26	0

Female Student and no Laptop = 4

Total Students = 62

Probability that a randomly selected student is a female and does NOT have a laptop = 4/62 = 0.0645

Probability that a randomly selected student is a female and does NOT have a laptop = 6.45%

2.5. Assume that the sample is a representative of the population of CMSU. Based on the data, answer the following question:

2.5.1 Find the probability that a randomly chosen student is a male or has a full-time employment

Employment	Full-Time	Part-Time	Unemployed
Gender			
Female	3	24	6
Male	7	19	3

Probability that a randomly chosen student is a male or has a full-time employment = 10/62 + 29/62 – 7/62 = 0.5161

P(Male U Full-Time Employment) = 51.6%

2.5.2 Find the conditional probability that given a female student is randomly chosen, she is majoring in international business or management.

Major	Accounting	CIS	Economics/Finance	International Business	Management	Other	Retailing/Marketing	Undecided
Gender								
Female	3	3	7	4	4	3	9	0
Male	4	1	4	2	6	4	5	3

Probability of female student randomly chosen, she is majoring in international business or management = (4/33) + (4/33) = 0.2424

P (International Business or Management| Female) = 24.24%

2.6 Construct a contingency table of Gender and Intent to Graduate at 2 levels (Yes/No). The Undecided students are not considered now and the table is a 2x2 table. Do you think graduate intention and being female are independent events?

Grad Intention	No	Yes
Gender		
Female	9	11
Male	3	17

P(Female ∩ Grad intention) = P(Female) * P(Grad Intention) - Condition of Independent events

P(Female) = 20/40 = 0.5

P(Grad Intention) = 28/40 = 0.7

P(Female) * P(Grad Intention) = 0.35

P(Female ∩ Grad intention = 0.275

Graduate intention and being female are not independent events

2.7 Note that there are four numerical (continuous) variables in the data set, GPA, Salary, Spending and Text Messages. Answer the following questions based on the data

2.7.1 If a student is chosen randomly, what is the probability that his/her GPA is less than 3?

GPA Gender	False	True
Female	25	8
Male	20	9

Above is the cross tab with the probability of his/her GPA is less than 3

student has GPA less than 3 = 17

Total Students = 62

Probability that his/her GPA is less than 3 = 17/62 = 0.2742

P (GPA < 3.0) = 27.42%

2.7.2 Find conditional probability that a randomly selected male earns 50 or more. Find conditional probability that a randomly selected female earns 50 or more.

Using contingency tables of Gender and Salary we get the total numbers of Male and Female and number of male and female earning 50 or more

Salary Gender	25.0	30.0	35.0	37.0	37.5	40.0	42.0	45.0	47.0	47.5	50.0	52.0	54.0	55.0	60.0	65.0	70.0	78.0	80.0
Female	0	5	1	0	1	5	1	1	0	1	5	0	0	5	5	0	1	1	1
Male	1	0	1	1	0	7	0	4	1	0	4	1	1	3	3	1	0	0	1

Salary Gender	False	True
Female	15	18
Male	15	14

Male earns more than or equal to 50 = 14

Total Male Students = 29

Conditional probability that a randomly selected male earns 50 or more = 14/29 = 0.4827

P (Salary >= 50|Male) = 48.27%

Female earns more than or equal to 50 = 18

Total Female Students = 33

Conditional probability that a randomly selected female earns 50 or more = 18/33 = 0.5454

P (Salary >= 50|Females) = 54.54%

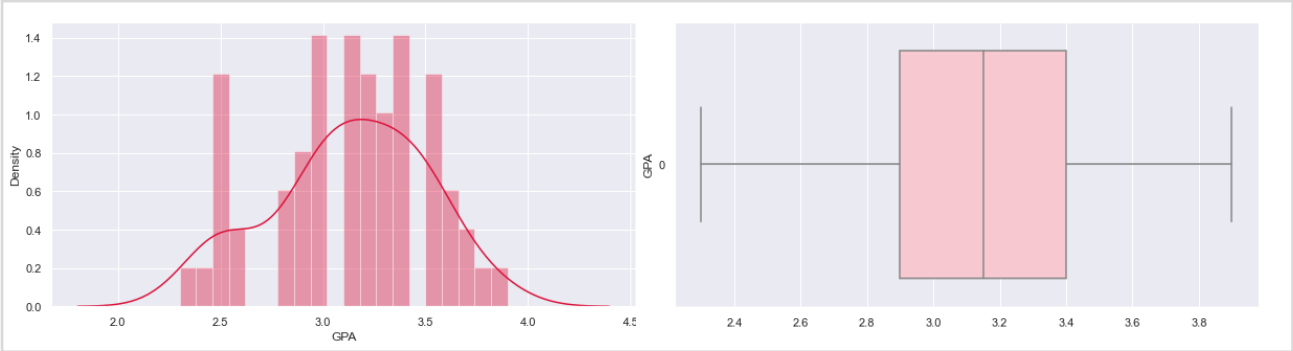
2.8.1 Note that there are four numerical (continuous) variables in the data set, GPA, Salary, Spending and Text Messages. For each of them comment whether they follow a normal distribution.

Considering each item one by one:

Ho: The data is normally distributed.

Ha: The data is not normally distributed.

1) GPA



From the above plots we can see that, GPA is Normally Distributed. Let us clarify using the Shapiro test as well.

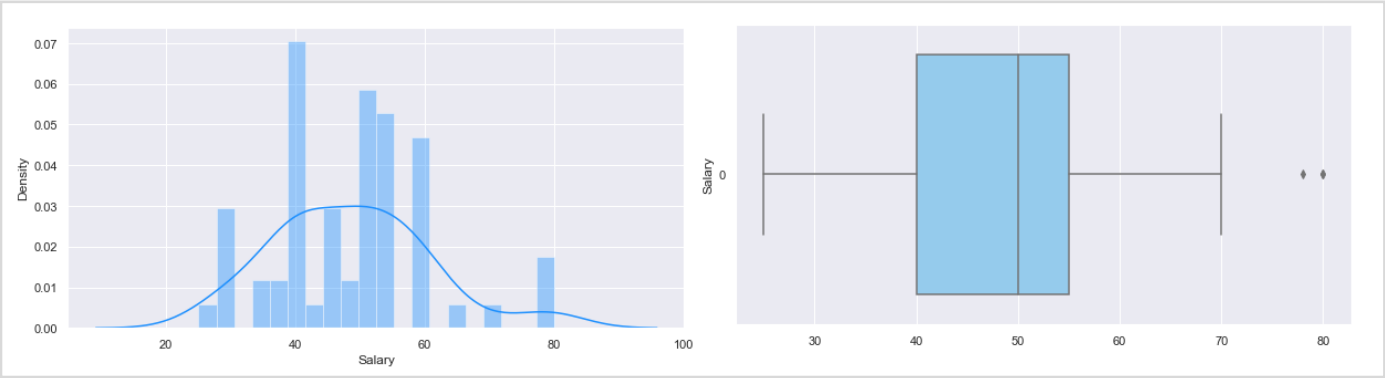
Shapiro-Wilk Test of Normality for GPA

- Test stat: 0.969
- P value: 0.112
  Since, P value > α (0.05), we fail to reject the Null hypothesis.

Hence, GPA is Normally distributed.

---

2) Salary



From the above plot, we can see that, Salary is slightly skewed towards right. Let us clarify using the Shapiro test as well.

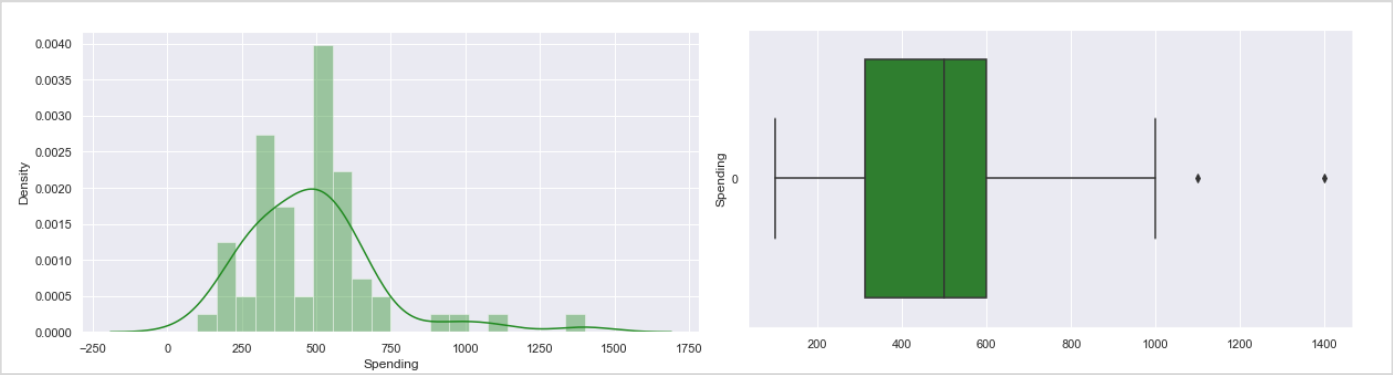
Shapiro-Wilk Test of Normality for Salary

- Test stat: 0.957
- P value: 0.028
  Since, P value < α (0.05), we reject the Null hypothesis.

Hence, Salary is NOT Normally distributed.

---

3) Spending



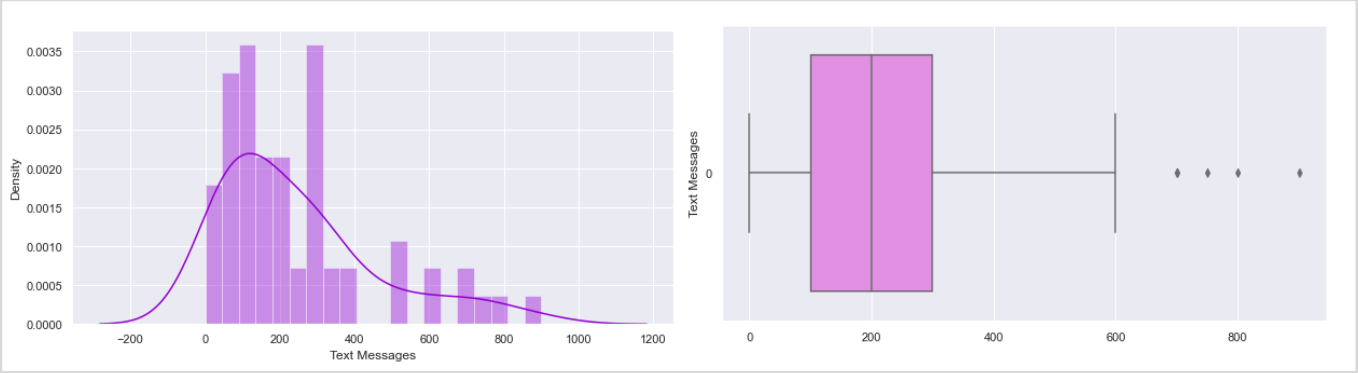
From the above plot, we can see that, Spending is skewed towards right. Let us clarify using the Shapiro test as well.

Shapiro-Wilk Test of Normality for Spending

- Test stat: 0.878
- P value: 1.68 x 10^{-5} Since, P value < α (0.05), we reject the Null hypothesis.

Hence, Spending is NOT Normally distributed.

4) Text Messages



From the above plot, we can see that, Text Messages is also skewed towards right. Let us clarify using the Shapiro test as well.

Shapiro-Wilk Test of Normality for Text Messages

- Test stat: 0.859
- P value: 4.324-06 Since, P value < α (0.05), we reject the Null hypothesis.

Hence, Text Messages is NOT Normally distributed.

2.8.2 Write a note summarizing your conclusions

Conclusion:

- Applied Displot and Box plot for four numerical (continuous) variables in the data set, GPA, Salary, Spending and Text Messages.
- We can see that, only GPA is normally distributed, which can be seen in the above plots that the whiskers of box plot of GPA are almost equal, where has the Salary, Spending, Text messages are slightly right skewed and also the length of whiskers are unequal and right skewed.
- The Shapiro wilk test which is used to test the normality, confirms the same.
- **GPA – Normal Distribution**
- **Salary, Spending, Text messages – Not normal Distribution**

Shingles Dataset

An important quality characteristic used by the manufacturers of ABC asphalt shingles is the amount of moisture the shingles contain when they are packaged. Customers may feel that they have purchased a product lacking in quality if they find moisture and wet shingles inside the packaging. In some cases, excessive moisture can cause the granules attached to the shingles for texture and colouring purposes to fall off the shingles resulting in appearance problems. To monitor the amount of moisture present, the company conducts moisture tests. A shingle is weighed and then dried. The shingle is then reweighed, and based on the amount of moisture taken out of the product, the pounds of moisture per 100 square feet are calculated. The company would like to show that the mean moisture content is less than 0.35 pounds per 100 square feet.

The file (A & B shingles.csv) includes 36 measurements (in pounds per 100 square feet) for A shingles and 31 for B shingles.

3.1 Do you think there is evidence that means moisture contents in both types of shingles are within the permissible limits? State your conclusions clearly showing all steps.

For this question, we will have to test dataset of shingles A and B separately to check if the moisture content in both are within permissible limits.

Step 1:

We need decide the Null and Alternative hypothesis

Ho : Mean moisture content <= 0.35 pound per 100 sq ft

Ha : Mean moisture content > 0.35 pound per 100 sq ft

Step 2:

Since the α is not given so,

Here we select α = 0.05

Step 3:

We do not know the population standard deviation and n = 36. So, we use 1 sample t test.

For Shingles A:

One sample t test

- t statistic: -1.473505
- P value: 0.07477633
  Since, P value > α (0.05), we fail reject the Null hypothesis.

So, the statistical decision is failing to reject the null hypothesis at 5% level of significance.

Conclusion

Hence, at 95% confidence level, there is sufficient evidence to prove that mean moisture content in A shingles is less than or equal to 0.35 pound per 100 square feet.

For Shingles B:

We do not know the population standard deviation and n = 31. So, we use 1 sample t test.

One sample t test

- t statistic: -3.10033
- P value: 0.002090
  Since, P value < α (0.05), we reject the Null hypothesis.

So, the statistical decision is to reject the null hypothesis at 5% level of significance.

Conclusion

Hence, at 95% confidence level, there is sufficient evidence to prove that mean moisture content in B shingles is more than 0.35 pound per 100 square feet.

3.2 Do you think that the population mean for shingles A and B are equal? Form the hypothesis and conduct the test of the hypothesis. What assumption do you need to check before the test for equality of means is performed?

For this, we will have to form a hypothesis and with the help of 2 sample t test find out whether both population means are same or not.

Step 1:

$H0$: μA - μB = 0 i.e. $\mu A = \mu B$

HA: μA - $\mu B \neq 0$ i.e. $\mu A \neq \mu B$

Step 2:

Since the α is not given so

Here we select α = 0.05.

Step 3:

T- Distribution

- t statistic: (1.2896282)
- P value: 0.20174965 Since, P value > α (0.05), we fail reject the Null hypothesis.

So, the statistical decision is failing to reject the null hypothesis at 5% level of significance.

Conclusion:

Hence, at 95% confidence level, there is sufficient evidence to prove that Population mean of shingles A is equal to Population mean of shingles B.

Assumptions needed to check before the test for equality of means is performed:

- The data values are continuous.
- The data values are independent.
- Measurement values for one observation do not affect measurement values of other observation.
- The data must be reasonably random.
- The data in each group is obtained via a random sample from the population.
- The data in each group is normally distributed.
- The variances for the two independent groups are equal.
